Towards virtual know ledge broker services for semantic integration of life science literature and data sources

Harrow, Ian ; Filsell, Wendy ; Woollard, Peter ; Dix, Ian ; Braxtenthaler, Michael ; Gedye, Richard ; Hoole, David ; Kidd, Richard ; Wilson, Jabe ; Rebholz-Schuhmann, Dietrich

Abstract: Research in the life sciences requires ready access to primary data, derived information and relevant knowledge from a multitude of sources. Integration and interoperability of such resources are critical for sharing content across research domains relevant to the life sciences. In this article we present a perspective review of data integration with emphasis on a semantics driven approach to data integration that pushes content into a shared infrastructure, reduces data redundancy and clarifies any inconsistencies. This enables much improved access to life science data from numerous primary sources. The Semantic Enrichment of the Scientific Literature (SESL) pilot project demonstrates feasibility for using already available open semantic web standards and technologies to integrate public and proprietary data resources, which span structured and unstructured content. This has been accomplished through a precompetitive consortium, which provides a cost effective approach for numerous stakeholders to work together to solve common problems.

DOI: https://doi.org/10.1016/j.drudis.2012.11.012

Accepted Manuscript

Title: Towards virtual know ledge broker services for semantic integration of life science literature and data sources

Authors: Ian Harrow, Wendy Filsell, Peter Woollard, Ian Dix, Michael Braxtenthaler, Richard Gedye, David Hoole, Richard Kidd, Jabe Wilson, Dietrich Rebholz-Schuhmann

PII: S1359-6446(12)00401-1
DOI: doi:10.1016/j.drudis.2012.11.012
Reference: DRUDIS 1122

To appear in:

Please cite this article as: Harrow, I., Filsell, W., Woollard, P., Dix, I., Braxtenthaler, M., Gedye, R., Hoole, D., Kidd, R., Wilson, J., Rebholz-Schuhmann, D., Towards virtual knowledge broker services for semantic integration of life science literature and data sources, *Drug Discovery Today* (2010), doi:10.1016/j.drudis.2012.11.012

Highlights

- Review of data integration in the Life Sciences with emphasis on semantic web.
- Concept of a Virtual Knowledge Broker (VKB) is applied to the Life Sciences.
- Semantic Enrichment of the Scientific Literature (SESL) demonstrator is an exemplar.
- Successful integration of data from numerous structured and unstructured sources.
- The Pistoia Alliance pre-competitive consortium delivered the pilot demonstrator.

Towards virtual knowledge broker services for semantic integration of life science literature and data sources

Ian Harrow[1,11], Wendy Filsell[2], Peter Woollard[3], Ian Dix[4], Michael Braxtenthaler[5,12], Richard Gedye[6,13], David Hoole[7], Richard Kidd[8], Jabe Wilson[9] and Dietrich Rebholz-Schuhmann[10]

[1]Ian Harrow Consulting (http://www.ianharrowconsulting.com)

[2]Unilever R&D, Colworth Science Park, Sharnbrook, Bedfordshire, MK44 1LQ UK

[3]GlaxoSmithKline, Medicines Research Centre, Gunnels Wood Road, Stevenage, Hertfordshire, SG1 2NY, UK

[4]AstraZeneca, 26T Mereside, Alderley Park, Cheshire, SK10 4TF UK

[5]Hoffmann-LaRoche, 340 Kingsland St, Nutley, NJ 07110-1199, USA

[6]Oxford University Press, Great Clarendon Street, Oxford, OX2 6DP, UK

[7]Nature Publishing Group, The Macmillan Building, 4 Crinan Street, London, N1 9XW, UK

[8]Royal Society of Chemistry, Thomas Graham House , Science Park, Milton Road, Cambridge, CB4 0WF, UK

[9]Reed Elsevier, 32 Jamestown Road, London, NW1 7BY, UK

[10] European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK

[11]Pfizer Worldwide Research and Development, Ramsgate Road, Sandwich, Kent, CT13 9NJ, UK

[12]Pistoia Alliance (http://www.pistoiaalliance.org)

[13] Richard Gedye Consulting (http://www.linkedin.com/in/richardgedye)

*Corresponding author***:** Harrow, I. (ianharrowconsulting@gmail.com)

Teaser:

Keywords:

[RE1]**Research in the life sciences requires ready access to primary data, derived information and relevant knowledge from a multitude of sources. Integration and interoperability of such resources are critical for sharing content across research domains relevant to the life sciences. In this article we present a perspective review of data integration with emphasis on a semantics driven approach to data integration that pushes content into a shared infrastructure, reduces data redundancy and clarifies any inconsistencies. This enables much improved access to life science data from numerous primary sources. The Semantic Enrichment of the Scientific Literature (SESL) pilot project demonstrates feasibility for using already available open semantic web standards and technologies to integrate public and proprietary data resources, which span structured and unstructured content. This has been accomplished through a precompetitive consortium, which provides a cost effective approach for numerous stakeholders to work together to solve common problems.**

In an era of information overload it is increasingly difficult for most researchers to find and access data to drive new insights and discoveries. It is essential that these researchers are able to query a wide variety of data sources through well-designed interfaces to gain ready access to all relevant data in the scientific literature, public and proprietary databases. In this article we review current approaches to data integration with particular emphasis on semantic web standards and technologies. These have been used to demonstrate technical feasibility for Virtual Knowledge Broker (VKB) services through the public SESL demonstrator (http://www.pistoia-sesl.org) which is focussed on human genes and uses the disease, Type 2 diabetes mellitus (T2DM), as the exemplar to test the premise [1]. This approach to data

integration promises to help researchers cope better with the data deluge and to create new business opportunities for the data providers.

The data deluge challenge

Modern life science research generates large volumes of experimental data by use of an ever-increasing number of technologies, especially those that operate in high throughput, such as biological screening, microarrays and next generation sequencing. These sources of structured data and accompanying metadata are also paralleled by enormous growth of unstructured data such as text found in the scientific literature. The researcher faces an enormous challenge in fully exploiting all these valuable resources to derive new insights and drive scientific discovery.

Hypotheses, conceptualised data and generalised facts from life science research are mainly delivered through the scientific literature comprising primary research articles, scientific reviews, conference proceedings, clinical data records, and patents. The evidence has been gathered from fundamental research such as transgenic experiments (e.g. gene knockouts), RNA and protein expression analyses, population genetics (e.g. GWAS), clinical studies, and crop studies [2]. Increasingly, scientific assertions are produced from the life science literature through automated text processing methods in combination with subsequent curation work [3,4].

Structured data from scientific databases and unstructured data from the scientific literature form the two pillars of scientific work. Primary research generates experimental evidence that is stored in numerous data repositories; for example, gene expression data resides in GEO/ArrayExpress [5] and information on the genetic causes of diseases can be found in the OMIM (Online Mendelian Inheritance in Man) database [6]. Such data are not always readily accessible within these resources, often individual data resources are incomplete and the delivered facts can be disconnected.

A shared data marketplace for life sciences

Retrieving only the pertinent data in the post-genomic era is very much like looking for needles in haystacks. Firstly the relevant data sources ('haystacks') have to be found and then each has to be searched, often separately, to locate the relevant data ('needles'). Surely, it would be far better if the user experience for the life science researcher was much more like that devised so successfully by internet businesses, such as online retail and travel or insurance comparison sites, which broker ready access to relevant information through a unified user interface.

The role of a knowledge broker has been described in the field of health care practitioner behaviour, which promotes interaction between researchers, patients and decision makers [7]. In this context, knowledge brokering enables mutual understanding of goals and culture, which also informs clinical policy and practise [8,9]. The knowledge broker role can also be found in the life sciences domain where an individual may provide a variety of expert analytical services, such as bioinformatics and text mining, and present the outcomes in a single aggregate report. Extending this approach to the user experience for unified search and retrieval, accessing numerous sources of data, describes the notion of a Virtual Knowledge Broker (VKB) [10].

In this review we argue that VKB services applied to life science data could ensure ready access to, and integration across, numerous primary data resources. It would also enable data providers to push their content to a virtual marketplace. Implementation of existing open standards should allow all data and information providers to shape their own content distribution, which could be delivered to a marketplace via a web portal in response to a researcher's queries. A scientist looking, for example, for information on genes causing disease in both the published literature and biological databases would not have to search each source separately, or compare and validate the retrieved results against each other, but would receive the aggregated information from the different available sources through a single interface designed for this purpose.

Data integration and Virtual Knowledge Broker services

Efforts to automatically process the scientific literature, such as PubMed abstracts and open access literature, have not yet led to significant repositories of facts, nor to the establishment of relevant and generally adopted data standards for the exchange of content between publishers and authors. To date, although the entire scientific research and publisher community can see real value in this approach, the integration of facts with data repositories and data distribution has not yet been achieved [11].

Five years ago, Louie and co-authors [12] described the challenges and opportunities of data integration for genomic medicine, and these remain the same today. Data marts (or warehouses) and the federation of data are the traditional approaches used to build large-scale infrastructures as illustrated in Figure 1. In the case of the mart, all data is integrated into a single infrastructure. This contrasts with federation, such as with caGrid, where scalable integration and interoperability is achieved through the harmonisation of separate database schemas [13]. These solutions are limited as it is very difficult to relocate or include additional types of data without significant development efforts. However, it is possible to design a more flexible infrastructure by utilising brokering services, leading to a solution that would even be open to shape workflows from external clients, such as the SADI services [14].

Open standards on the web, such as the Resource Description Framework (RDF) and the Web Ontology Language (OWL), enable data integration of distributed resources with encoded meaning (i.e. semantics). Semantic web standards and technologies are well suited to meet the challenges of data integration, as evidenced by a growing number of bioinformatics resources where data is distributed in an open infrastructure, such as the Linked Life Data prototype [15], Chem2Bio2RDF [16], Neurocommons [17] and the SESL public demonstrator, which is summarised in the next section.

The integration of structured and unstructured data into an open infrastructure requires shared data standards. Several open standards for better semantic support have been proposed, such as the use of open access terminologies and ontologies [11], as well as the application of data exchange formats such as RDF and OWL [18]. These enable communication across the web and support both sharing and exploitation of data resources. The Link Open Drug Data (LODD) task force within the World Wide Web Consortium (W3C) is a good example of an emerging technology for linking data sets to enable data integration that support research in academia and industry [19].

Thinking beyond data integration, semantic interoperability and distributed exploitation of public data, numerous requirements have to be met to support the life science research community. First, public and proprietary data have to be fully accessible in a shared and seamless manner, including the full text of the scientific literature. Second, different kinds of numerical data with semantic annotations, such as experimental metadata, must be included and even, ideally, extended to supplementary data, tables and figures from the scientific publications. Third, the semantic facts have to be integrated into and embedded in the shared infrastructure. The SESL demonstrator, summarised in the next section, shows that it is feasible to satisfy these requirements and to achieve semantic integration of literature and data resources through the implementation of VKB services.

Towards Virtual Knowledge Broker services

Unstructured data sources: full text scientific literature

The participating data provider companies (Elsevier, Nature Publishing Group, Oxford University Press, Royal Society of Chemistry) contributed 638,088 scientific full text publications to the SESL project and a further 232,665 full text documents were sourced from Europe PubMed Central. These two sources formed the literature corpus used to develop the brokering framework. The publicly accessible SESL demonstrator contains a subset of the

literature corpus used for its development, comprising 20,168 full text publications, which were released to the public domain by the data providers for this purpose.

All full text documents in the literature corpus were processed in a similar manner. The identification of gene and protein names was achieved using the terminological resource, LexEBI, which served as a term repository for the biomedical domain and provides references to the entries from the different primary data providers [20]. Identification of diseases was achieved using UMLS terminological resource [21]. All sentences containing a pair composed of a gene and a disease were identified and loaded into a triple store database. Sentence provenance was retained as part of the process, that is, the reference to the sentence, paragraph and the document source, the digital object identifier (DOI) and associated publication reference data.

Structured data sources: omics databases

The current UniProtKB triple store delivers an integrated representation of the database based on RDF triples [22]. The human subset of this data source used to build the SESL demonstrator, contains 20,272 proteins, 100,723 functional annotations and 13,897 protein–protein interactions, all rendered as triples. Data relevant to T2DM was imported from Gene eXpression Atlas (GXA) resulting in the integration of data from 138 experiments [5]. Disease annotations from the Experimental Factor Ontology (EFO) used in the GXA data repository were normalised to the Disease Ontology (DO) so the existing mappings to UMLS could be used. Gene expression data was also loaded for the pancreas because this organ has a major role in blood sugar regulation, which is fundamental to diabetic diseases. Genetic disease data was imported from OMIM Morbid Map and the terminologies were normalised to UMLS, which was used as the disease terminology source.

Prototype Virtual Knowledge Broker services

The prototype SESL demonstrator shows how an open semantic web infrastructure can integrate different primary data sources as VKB services. This approach makes it

straightforward to ensure that any duplicated data from two distinct primary data resources displays [RE2]unique information in the graphical user interface of the demonstrator. The VKB layer resolves duplicate data from the multiple primary data sources. It extracts assertions and metadata from the primary sources and transforms the extracted data into RDF triples. Following semantic normalisation, the processed data is accessible from a RDF triple store through a SPARQL endpoint [23–25]. The architecture is described in Figure 2.

The integration of content from different information providers is particularly beneficial as they often render different interpretations from the same data resources (Figure 3). For example, the three genes *LIPC*, *SLC30A8* and *UCP1* are linked to diabetes according to UniProt, but the disease stems from the 'polymorphisms' field rather than the field named 'Involvement in disease'. By applying existing open data standards in a consistent manner, it is possible to export all the relevant data into a single infrastructure.

The SESL demonstrator brokers static data resources in RDF that stem from selected public bioinformatics databases and the publication corpus, as described above. A set of triple stores comprising the broker system have been tested for their flexibility and extensibility. It has proven possible to distribute the brokered data services over different compute engines in a federated grid cluster. This approach offers flexibility so that static RDF resources can be redistributed freely, loaded into any decentralised location and also be kept in traditional relational databases.

The SESL graphical user interface, illustrated in Figure 4, shows how a single query, for example, for a gene name, returns an aggregated set of gene and disease relationships, where the results are derived from multiple primary data sources [26]. Therefore, the SESL demonstrator shows that it is technically feasible to deliver semantic integration of primary data sources through VKB services. This approach has the potential to simplify and improve the user experience and to more fully explore all the information available about entities such as genes associated with a disease, including any conflicting assertions.

The role of semantic web standards

Semantic web technologies and standards are particularly well suited to accomplish data aggregation and integration [27,28], but require full implementation of open semantic web standards to realise the potential to reduce data redundancy and improve consistency across the disparate sources.

Standardising the assertions from the scientific literature requires the reuse of public semantic resources such as UMLS [21] and BioLexicon/LexEBI [29]. It is also important to use existing metadata standards, such as IeXML [30], SKOS [31] and Dublin Core [32]. This enables interoperability and seamless integration of all literature content with other data resources while lowering the overhead costs for any literature providers to participate. Other relevant standards that should be considered include MIBBI [33] and the Open Archive Initiative Protocol for Metadata Harvesting [34]. For the SESL demonstrator, the extracted assertions have been represented as RDF triples and used as a minimum information entity from the scientific literature [33]. Such assertions can be formally combined with provenance to give 'nanopublications' which enable microattribution [35,36]. Nanopublications rather than narrative full text articles have also been proposed as an alternative central information unit for future scholarly communication [37].

Standards are already available to achieve semantic interoperability of distributed and redundant data repositories using semantic web technology, and scientific literature providers can make use of these existing standards to be compliant. However, access to and processing of the literature through a text mining service is still necessary to deliver the assertions that support the brokering approach. An increasing number of publishers and data providers are already getting involved in activities, either in-house or working together in precompetitive consortia, to move towards data and literature content being colocated or associated in such a manner that they can be exploited using the semantic web [22,39,40,41].

It is conceivable that analytical software tools on the web could identify suitably tagged data stored in suitable primary repositories with minimal external influence. Such semantically tagged data resources could self-register for sharing or licensing conditions before consumption by analytical web services that produce and consume RDF as automated workflows. Examples of such open semantic frameworks are the SSWAP (Simple Semantic Web Architecture and Protocol), SADI/SHARE and S3DB semantic web frameworks [14,38]. SHARE exposes SADI web services as if they were a virtual, distributed SPARQL endpoint to achieve semantic integration.

Sharing precompetitive competencies

The SESL pilot project set out to evaluate the feasibility of using open semantic web standards to build a knowledge brokering system for life science data. It was commissioned by the Pistoia Alliance, which is a precompetitive alliance of life science organisations and institutions. The SESL project team comprised of representatives from five large companies engaged in life science research (Pfizer, GlaxoSmithKline, AstraZeneca, Unilever and Hoffmann-LaRoche) three scholarly publishers (Oxford University Press, Nature Publishing Group and Elsevier), one learned society (Royal Society of Chemistry) and an academic partner (EMBL-EBI). Each representative brought different expertise and views to the team, with all recognising a common set of challenges, these include:

(i) the large volume and complexity of life science data and published literature is now beyond the ability of a single user or organisation to query or manage in a comprehensive and cost effective manner;

(ii) open semantic web standards need to be supported and promoted to encourage their widespread adoption;

(iii) data providers are the experts in the technologies required for management and integration of their data;

(iv) users need to be able to readily access and exploit all available data in a timely manner;

(v) value can be added to data by effective aggregation of many primary data sources.

By working together in a precompetitive manner we have shown that existing open semantic web standards and technologies are sufficient to integrate data from numerous primary data sources. This will bring benefit to both data consumers and providers working in life sciences.

Concluding remarks

In this perspective we have discussed the challenges and opportunities raised by the growing deluge of data being generated in life sciences. We have placed particular emphasis on semantics-driven approaches to improve data integration and access through VKB services. The SESL project has developed a public demonstrator that, for the first time, shows that existing open semantic web standards and technologies are sufficient to integrate structured and unstructured data derived from a selection of public and proprietary data sources. This fully functional prototype has been developed over a period of approximately one year and on a modest budget. It shows how a precompetitive consortium, comprising of members from different parts of the scientific community can share costs and risks to demonstrate technical feasibility for data integration through VKB services.

Looking to the future, principles similar to those used in the SESL project are being applied by the Open PHACTS (Open PhArmacological Concepts Triple Store) consortium, which is funded by the Innovative Medicines Initiative (IMI) for three years. OpenPHACTS brings together academic and pharmaceutical partners to design and implement an open source, open standards and open access innovation platform, the Open Pharmacological Space (OPS) that is designed to deliver semantic interoperability for drug discovery [41].

Acknowledgements

This topic was first discussed at an EMBL EBI Industry Programme workshop on semantic enhancement of the scientific literature (SESL). The subsequent SESL project was supported and funded by the Pistoia Alliance, which is a precompetitive and not-for-profit organisation. We thank the following people for valuable contributions, which provided substrate for this perspective: Christoph Grabmuller , Silvestras Kavaliauskas , Misha Kapushesky, Jennifer Cham, Dominic Clark, Johanna McEntyre (EMBL-EBI), Ashley George, Nick Lynch, John Wise (Pistoia Alliance) Catherine Marshall, Nigel Wilkinson (Pfizer), Claire Bird, Richard O'Beirne (OUP), Ian Stott (Unilever) and Mike Westaway (AstraZeneca).

Conflicting Interests

The authors declare that they have no conflicting interests.

References

1. McCarthy, M. I. (2011) The importance of global studies of the genetics of type 2 diabetes. *Diabetes Metab. J.* 35, 91–100

2. McCarthy, M. I. and Zeggini, E. (2009) Genome-wide association studies in type 2 diabetes. *Curr. Diab. Rep.* 9, 164–171

3. Kirsch, H. *et al.* (2006) Distributed modules for text annotation and ie applied to the biomedical domain. *Int. J. Med. Inform.* 75, 496–500

4. Rebholz-Schuhmann, D. *et al.* (2005) Facts from text–is text mining ready to deliver? *PLoS Biol*. 3, E65

5. Parkinson, H. *et al.* (2011) ArrayExpress update–an archive of microarray and high-throughput sequencing based functional genomics experiments. *Nucleic Acids Res.* 39, D1002–D1004

6. McKusick, V. (2007) Mendelian Inheritance in Man and Its Online Version, OMIM. *Am. J. Hum. Genet.* 80, 588–604

7. Dobbins, M. *et al.* (2009) A description of a knowledge broker role implemented as part of a randomized controlled trial evaluating three knowledge translation strategies. *Implement. Sci.* 4, 23–32

8. Ward, V. *et al.* (2009) Knowledge brokering: Exploring the process of transferring knowledge into action. *BMC Health Services Research* 9, 12–18

9. Meyer, M. (2010) The Rise of the Knowledge Broker. *Science Communication* 32, 1180127

10. Verona, G. *et al.* (2007) Innovation and virtual environments: Towards virtual knowledge brokers. *Organization Studies.* 27, 765–788

11. Harland, L. *et al.* (2011) Empowering industrial research with shared biomedical vocabularies. *Drug Discov. Today* 16, 940–947

12. Louie, B. *et al.* (2007) Data integration and genomic medicine. *J. Biomed. Informatics.* 40, 5–16

13. Oster, S. *et al.* (2008) caGrid 1.0: an enterprise Grid infrastructure for biomedical research. *J. Am. Med. Inform. Assoc.* 15, 138–149

14. Wilkinson, M. D. *et al.* (2011) The Semantic Automated Discovery and Integration (SADI) Web service Design-Pattern, API and Reference implementation. *J. Biomed. Semantics* 2, 8–31

15. Linked life Data prototype: a semantic data integration platform for the biomedical domain

16. Chen, B. *et al.* (2010) Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics* 11, 255

17. Ruttenberg, A. *et al.* (2009) life sciences on the Semantic Web: the Neurocommons and beyond. *Briefings in Bioinformatics.* 10, 193–204

18. Chen, B. *et al.* (2012) Improving integrative searching of systems chemical biology data using semantic annotation. *J. Chemoinform.* 4, 6–17

19. Samwald, M. *et al.* (2011) Linked open drug data for pharmaceutical research and development. *J. Chemoinform*. 3, 19–25

20. Spasic, I. *et al.* (2008) Facilitating the development of controlled vocabularies for metabolomics technologies with text mining. *BMC bioinformatics* 9, S5

21. Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 32, D267–D270

22. Apweiler, R. *et al.* (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.* 39, 214–219

23. W3C: SPARQL Protocol for RDF http://www.w3.org/tr/rdf-sparql-protocol

24. W3C: SPARQL Query Language for RDF http://www.w3.org/tr/rdf-sparql-query

25. Resource Description Framework (RDF) http://www.w3.org/rdf

[RE3]26. Kim, J. J. and Rebholz-Schuhmann, D. (2008) Categorization of services for seeking information in biomedical literature: a typology for improvement of practice. *Briefings in Bioinformatics* 9, 452–465

27. Stephens, S. *et al.* (2006) Aggregation of bioinformatics data using Semantic Web technology. *J. Web Semantics.* 4, 216–221

28. Goble, C. and Stevens, R. (2008) State of the nation in data integration for bioinformatics. *J. Biomed. Informatics* 41, 687–693

29. Sasaki, Y. *et al.* (2008) Biolexicon: a lexical resource for the biology domain. *Proc. of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008).* 3, 109–116

30. Rebholz-Schuhmann, D. *et al.* (2006) IeXML: towards a framework for interoperability of text processing modules to improve annotation of semantic types in biomedical text. *BioLINK, ISMB 2006, Fortaleza, Brazil*

31. Simple Knowledge Organization System Reference http://www.w3.org/TR/skos-reference

32. Dublin Core Metadata Initiative: Metadata Terms
http://dublincore.org/documents/2010/10/11/dcmi-terms

33. Taylor, C. F. *et al.* (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.* 26, 889–896

34. Open Archives Initiative Protocol for Metadata Harvesting
http://www.openarchives.org/pmh

35. Patrinos, G. *et al.* (2012) Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain. *Human Mutation.* 33, 1503-12

36. Nonapub.org: Learn how to create, find, use and cite nanopublications
http://www.nanopub.org

37. Mons, B. *et al.* (2011) The Value of Data. *Nat. Genet*. 43, 281–283

38. Almeida, J. S. *et al.* (2010). S3DB core: a framework for RDF generation and management in bioinformatics infrastructures. *BMC Bioinformatics* 11, 387

39. Attwood, T.K. *et al.* (2009) Calling International Rescue: knowledge lost in literature and data landslide! *Biochemical Journal* 424 317–333

40. Shotton, D. *et al.* (2009) Adventures in semantic publishing: exemplar semantic enhancements of a research article. *PLoS Comput. Biol*. 5, E1000361

41. Williams, A.J. *et al.* (2012) Open PHACTS: semantic interoperability for drug discovery. *Drug Discov. Today.* 5, 16–27

Figure 1: Different data integration solutions. [RE4]

Comparing data integration of different kinds: mart (left), federation (middle) and brokering (right). In the mart, all data is integrated into a complex schema, whereas federation requires that data consistency is achieved through interoperable data export and import interfaces. Brokering requires a Virtual Knowledge Broker (VKB), which uses open standards to expose

the data and compliance with provenance and licensing (P&L) to be able to integrate the content from the primary data sources into the user-oriented data repository.

Figure 2: Architecture of the Virtual Knowledge Broker (VKB) service.

The VKB service layer comprises of assertion and metadata extraction from primary sources, business rules to determine data extraction, transformation of extracted data to RDF triples, public vocabularies, the triple store for integration and aggregation and the SPARQL endpoint.

Abbreviations: OMIM: Online Mendelian Inheritance in Man; RDF: Resource Description Framework.

Figure 3: Minimal configuration to test technical feasibility.

The minimal configuration to test for de-duplication of data in a virtual knowledge brokering service is shown conceptually, where two triple stores have identical structure, but primary source content can overlap. UniProtKB content was used to test this condition.

Abbreviations: OMIM: Online Mendelian Inheritance in Man; NPG: Nature Publishing Group; OUP: Oxford University Press; RSC: Royal Society of Chemistry.

Figure 4: The SESL public demonstrator for Virtual Knowledge Broker (VKB) services.

The simple graphical user interface (GUI) is shown schematically to illustrate an exemplar VKB service where a single query by gene and/or disease can return a single set of aggregated results for gene and disease relationships derived from numerous primary data sources.

Abbreviations: OMIM: Online Mendelian Inheritance in Man; SESL: Semantic Enrichment of the Scientific Literature.

Figure 2: Different data integration solutions

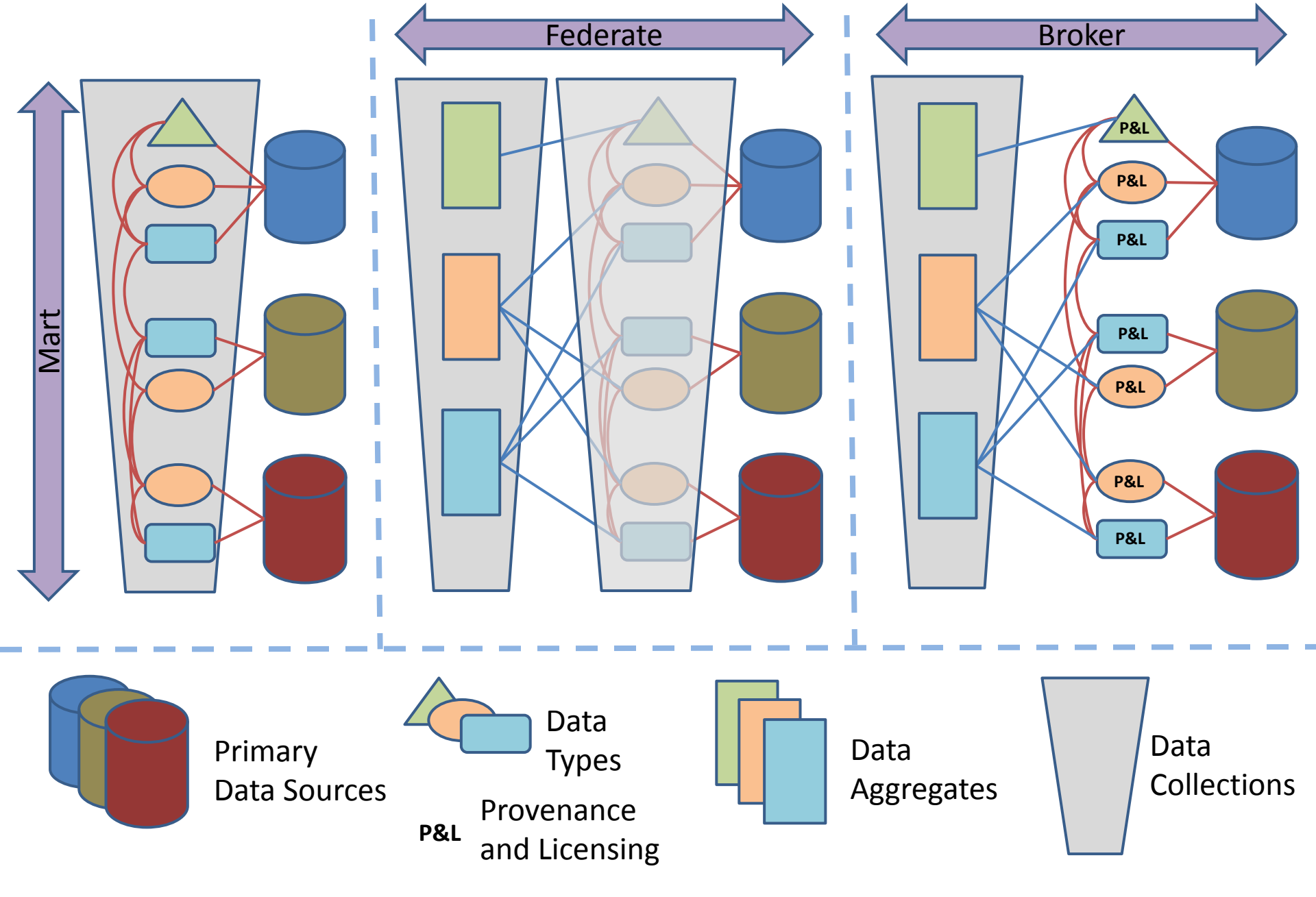

Figure 2: Architecture of the Virtual Knowledge Broker (VKB) Service

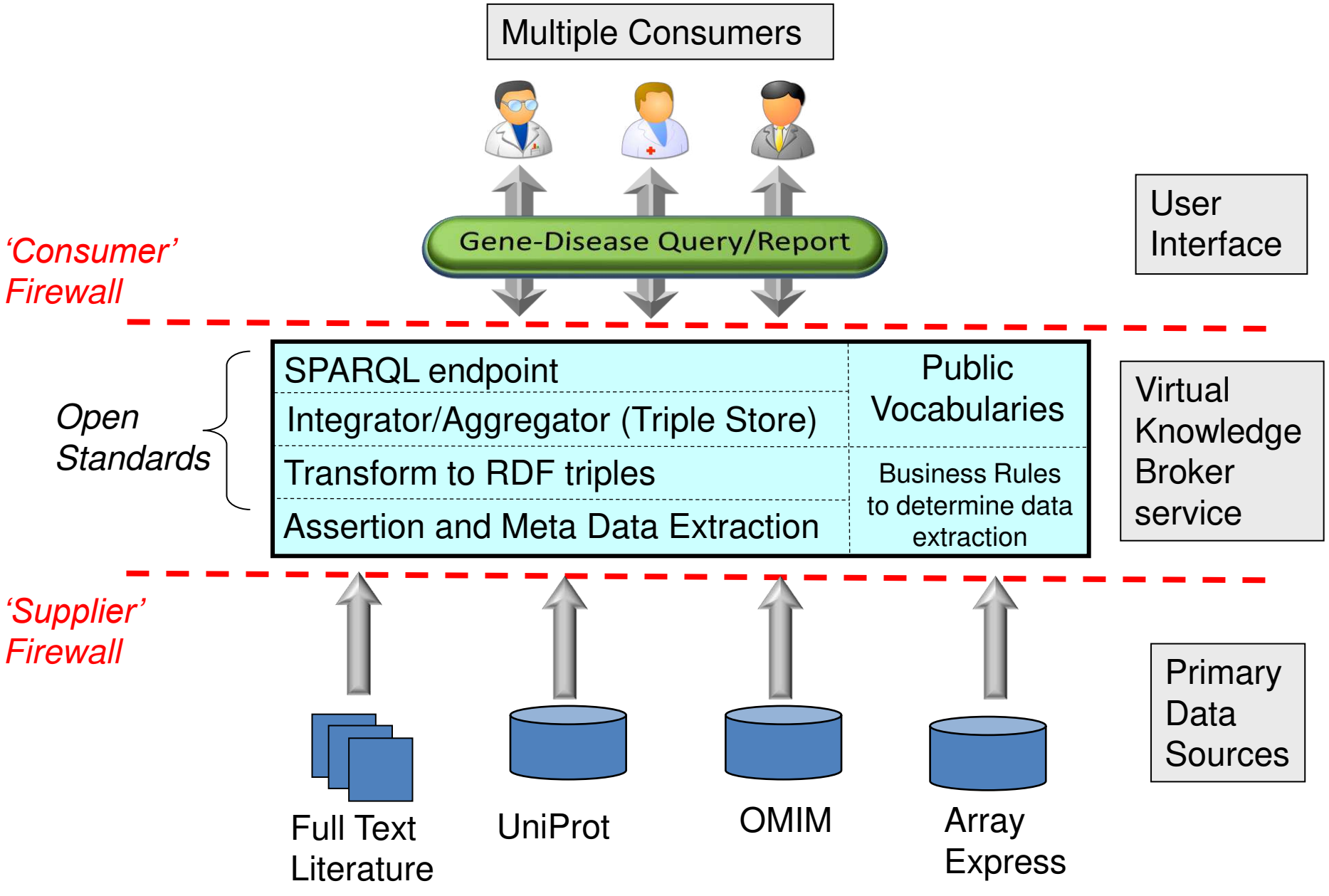

Figure 3: Minimal configuration to test technical feasibility

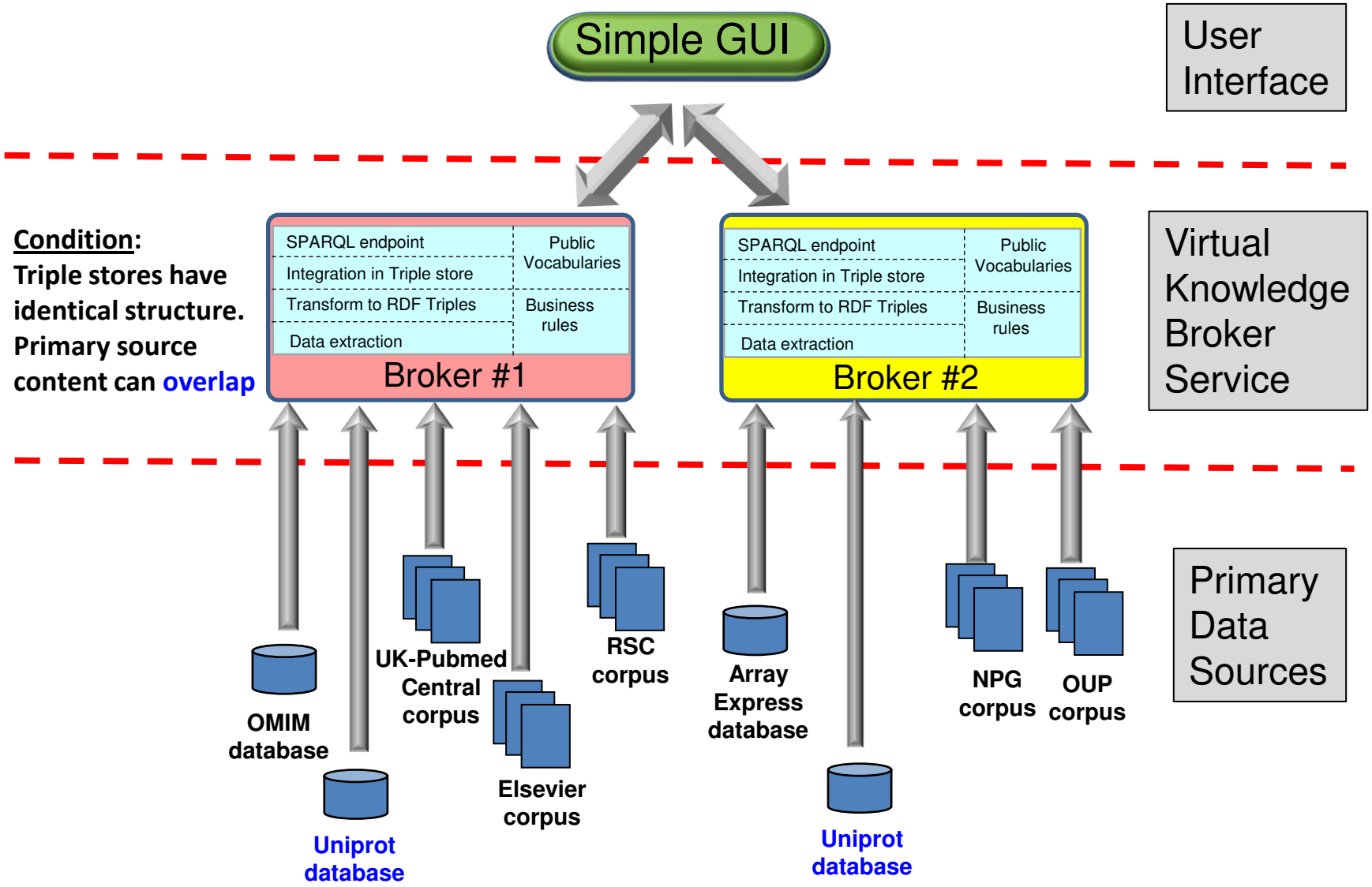

Figure 4: The SESL public demonstrator for Virtual Knowledge (VKB) services